Beegle: from literature Mining to Disease Gene Discovery

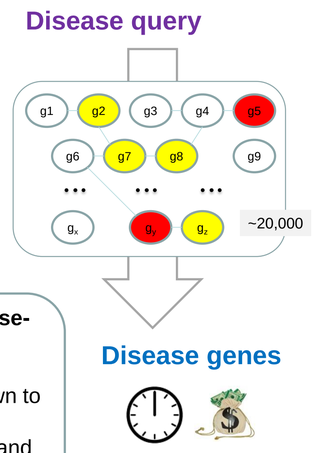
Sarah ElShal[1,2], Léon-Charles Tranchevent[1, 2, 3, 4, 5], Jesse Davis[6], and Yves Moreau[1,2]

WHY: Disease Gene Discovery

Determining which genes cause which diseases is an important yet challenging problem. It has a variety of applications that range from **DNA screening** and **early diagnosis**, to gene **sequence analysis** and **drug development**.

However, it is **resource intensive** both in terms of **time** investment and monetary **cost**. **Traditionally**, disease-gene identification is approached **manually** and is conducted in two phases:

1. **narrow down a large set of candidate genes** (*e.g.*, the whole genome) into a significantly smaller set of genes that has a high probability of containing a disease causing gene (*e.g. linkage analysis, genome sequencing, and association studies*).
2. **evaluate the selected genes** to confirm which of those candidates are **truly disease causing** (through wetlab experimentation)



In this work we present ***Beegle***, an online search and discovery engine for **disease-gene prioritization** that **entirely automates** the first phase of disease-gene discovery.

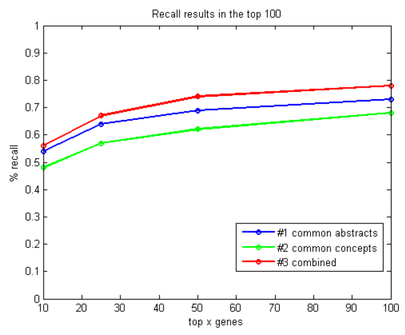
- Its starts by **mining the literature** to automatically extract a set of genes known to be linked with a given query (the search phase).
- Then it integrates multiple sources of **genomic information** to learn a model and rank a set of candidate genes (*e.g.*, the human genome) according to a selection of the output list of known genes retrieved in the first phase (the discovery phase).

benchmark: OMIM validation sets

We evaluate our tool based according to a **validation set** that we extracted from the disease-gene associations that are recorded in the **OMIM catalog**. Hence, for every OMIM disease, we have the list of genes associated with it.

For the **search phase**:

- We evaluate **how well *Beegle* identifies known** disease-gene associations.
- For a given OMIM disease, we measure how many of its associated genes are returned at the top ranks.
- using the **best rank**, we achieve the best recall of 56% in the top 10 and **78% in the top 100** returned genes.
- using co-occurrence alone results in an average recall of 54% in the top 10 vs. 73% in the top 100.
- using concept profile similarity alone results in an average recall of 48% in the top 10 vs. 68% in the top 100.



For the **discovery phase**:

- We evaluate the **suitability of the returned genes** in the search phase **to serve as input to train genomic models** and generate novel hypothesis.
- we employ an **evaluation methodology that mimics real discovery** by using rolled-back data to generate the gene prioritizations, and then by testing on disease-gene associations that were reported after the training data was collected.
- For a given OMIM disease, we once use its associated genes (as recorded in the OMIM database) as the input to train its model, and once we use the top 10 genes returned by *Beegle*. Then we prioritized the rest of the human genome for novel hypothesis.
- We compare the resulting prioritizations from both models, and measure how much we recall from the test associations in the top ranked genes.
- We observe comparable true positive rates given the two inputs in the top 5%, 10%, and 30%.
- We also used **another literature-based benchmark** where the input is manually-extracted from the literature. We observe that **using the top-10** genes returned by *Beegle* (in the search phase) **improves the TPR** by 44%, 27%, and 9% in at the top 5%, 10%, and 30% prioritized genes respectively.

Input set	TPR in top 5%	TPR in top 10%	TPR in top 30%
OMIM-reported	35%	45%	67%
***Beegle's* top-10**	37%	46%	67%

Input set	TPR in top 5%	TPR in top 10%	TPR in top 30%
Manually-extracted	28.6%	38.1%	71.4%
Beegle's top-10	**41.2%**	**48.5%**	**77.5%**

discussion: WHY *Beegle*?

1. *Beegle* applies a **novel combination of text mining approaches** that proves to **work better** than standalone approaches (which are applied in isolation in other tools, *e.g. MeSHOP*)
2. *Beegle* is able to **automatically search** the literature for known disease-gene associations (e.g. OMIM associations)
3. *Beegle* **automatically generates an interesting training set** to build models for novel genes prediction.
4. The web interface is **user-friendly** (especially with the online tutorial available on the home page). That is in addition to the literature evidence it attaches with every ranked gene.

WHAT: online search and discovery engine

available at: http://beegle.esat.kuleuven.be/

Steps on *Beegle*
Nice features

0- The user enters the query of interest in the search bar (which is associated with an auto-complete option)

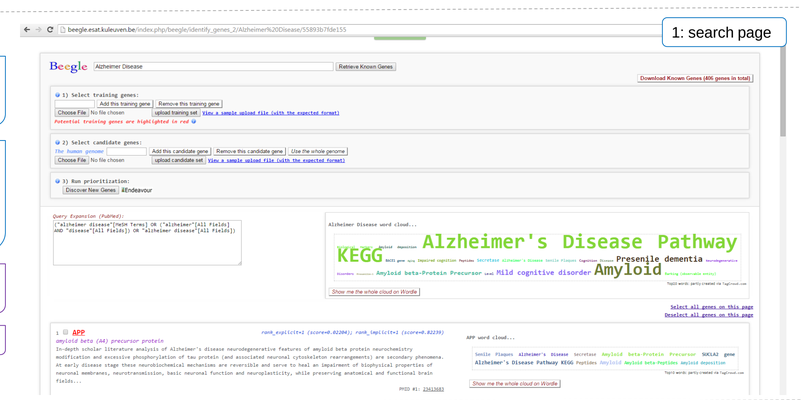
*A google-like search bar that accepts **any PubMed query***

1- *Beegle* runs the **search phase** and analyzes the literature in order to present back an ordered list of the **genes that are already known** to be linked with the given query.

- In order to proceed with the discovery phase, the user **defines a training set** according to the presented output list or by manual selection.
- The user also **defines the candidate set** to be prioritized either manually or by selecting the whole genome.

Attaches literature evidence (1-3 common abstract snippets + common concepts)

Highlights potential training genes

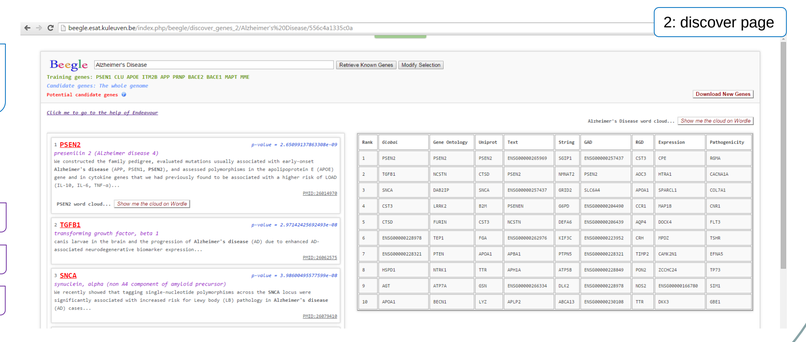


2- *Beegle* runs the **discovery phase** (through Endeavour) and builds training models in order to **prioritize the candidate set** for novel gene hypothesis.

*Attaches **recent common publications** (if exists)*

Attaches Endeavour (separate genomic) rankings

Highlights potential candidate genes

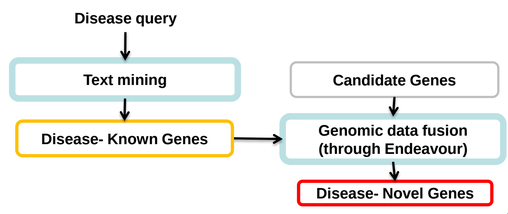


The results are based on the 2013 release of literature and other genomic data

HOW: generating the associations

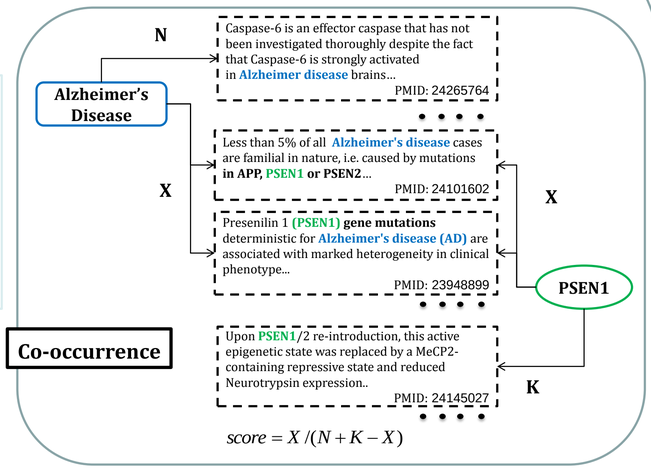
The pipeline

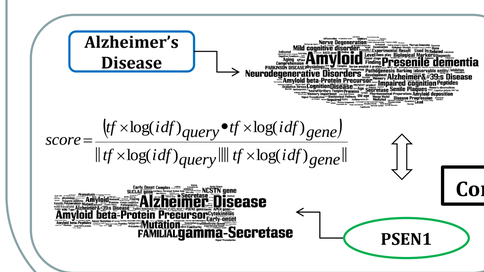
Beegle proceeds in **two phases**. First, based on the user query, it automatically analyses the literature to identify the genes that are potentially related to the query. Second, it uses these genes (identified in the first step) as a seed set that is provided to *Endeavour*, which then analyses a number of genomic data sources to produce a final prioritization of the candidate genes.



The search phase

Beegle applies **two text mining approaches** to identify the genes most related to a given disease. The first one is based on the number of abstracts in which the disease and a given gene **co-occur**. The second approach is based on the number of **common concepts** between the abstracts linked to a gene and the given disease.
Beegle assigns a final gene-disease score by **combining the output of both approaches**, which corresponds to a meta-analysis procedure using **the best rank** calculated by each approach.

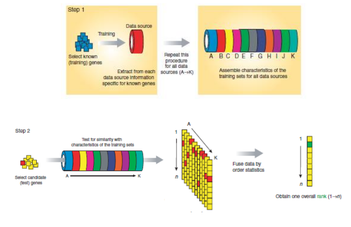




Best Rank $rank = \min(R1, R2)$

The discovery phase

- *Beegle* integrates the methodology of *Endeavour* to generate the final gene prioritization for a given disease. *Endeavour* relies on three inputs: (1) a set of training genes known to be linked to the disease or query under study, (2) a set of data sources that are used to build the disease models using the training genes, and (3) a set of candidate genes to investigate (*i.e.*, to prioritize). Per data source, ***Endeavour* ranks the candidate genes** according to how similar a gene is to the corresponding model, therefore providing one ranked list for each data source. To combine the lists, *Endeavour* applies order statistics to produce a single ranking, which is the final prioritization list for the given disease.
- *Beegle* **uses a user-selected set of the top disease-genes returned in the search phase as the training set for *Endeavour***. Then it runs *Endeavour* to generate the final disease-gene prioritization.



[1]KU Leuven, Department of Electrical Engineering (ESAT) STADIUS, 3001 Leuven, *Belgium* [2]iMinds Future Health Department, Leuven, *Belgium*
[3]Inserm UMR-S1052, CNRS UMR5286, Cancer Research Centre of Lyon, Lyon, France [4]Université de Lyon 1, Villeurbanne, France
[5]Centre Léon Bérard, Lyon, France [6]KU Leuven, Department of Computer Science (DTAI), 3001 Leuven, *Belgium*

sarah.elshal@esat.kuleuven.be
yves.moreau@esat.kuleuven.be